MBE Advance Access published December 4, 2015

Article: Discoveries

Long distance dispersal shaped patterns of human genetic diversity in Eurasia

Isabel Alves[1,2,3], Miguel Arenas[1,2,4], Mathias Currat[5], Anna Sramkova Hanulova[1,2], Vitor C. Sousa[1,2], Nicolas Ray[6], Laurent Excoffier[1,2]

[1)] Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution, University of Bern, Berne, Switzerland.

[2)] Swiss Institute of Bioinformatics, Lausanne, Switzerland.

[3)] Population and Conservation Genetics Group, Instituto Gulbenkian de Ciência, Oeiras, Portugal.

[4)] Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal.

[5)] Anthropology, Genetics and Peopling History Lab, Anthropology Unit – Department of Genetics & Evolution - Anthropology Unit, University of Geneva, Geneva, Switzerland.

[6)] EnviroSPACE Lab., Institute for Environmental Sciences, University of Geneva, Switzerland.

Correspondence: Isabel Alves, isabel.alves@iee.unibe.ch; Laurent Excoffier, laurent.excoffier@iee.unibe.ch

Running title: Long distance dispersal shaped human diversity

Abstract

Most previous attempts at reconstructing the past history of human populations did not explicitly take geography into account, or considered very simple scenarios of migration and ignored environmental information. However, it is likely that the Last Glacial Maximum (LGM) affected the demography and the range of many species, including our own. Moreover, long-distance dispersal (LDD) may have been an important component of human migrations, allowing fast colonization of new territories and preserving high levels of genetic diversity. Here, we use a high-quality microsatellite dataset genotyped in 22 populations to estimate the posterior probabilities of several scenarios for the settlement of the Old World by modern humans. We considered models ranging from a simple spatial expansion to others including LDD and a LGM-induced range contraction, as well as Neolithic demographic expansions. We find that scenarios with LDD are much better supported by data than models without LDD. Nevertheless, we show evidence that LDD events to empty habitats were strongly prevented during the settlement of Eurasia. This unexpected absence of LDD ahead of the colonization wave front could have been caused by an Allee effect, either due to intrinsic causes such as an inbreeding depression built during the expansion, or to extrinsic causes such as direct competition with archaic humans. Overall, our results suggest only a relatively limited effect of the LGM-contraction on current patterns of human diversity. This is in clear contrast with the major role of LDD migrations, which have potentially contributed to the intermingled genetic structure of Eurasian populations.

Introduction

Archaeological and genetic evidence suggest that anatomically modern humans (AMH) originated in Africa less than 200 thousand years ago (Kya) giving rise to all extant world populations through a major migration wave that left Africa around 60-70 Kya (Mellars 2006a, b). Recently, whole-genome analyses of ancient DNA have also revealed that AMH admixed with archaic humans during their expansion out of Africa leaving a contribution of 1 to 6% in the genomes of extant Eurasians (Green, et al. 2010; Reich, et al. 2010; Reich, et al. 2011). Although the evolutionary history of our own species is well known (Henn, et al. 2012), many of its aspects like the number of colonization waves out of Africa, the exact migration routes followed by modern humans, or the effect of cultural developments on the genetic exchange and the inhabitable range of our species, remain poorly understood. With the recent development of modeling techniques the importance of the environment in shaping the distribution of human populations and, therefore, in conditioning the dispersal of our own species has been increasingly recognized (Banks, et al. 2008; Eriksson, et al. 2012; Ray, et al. 2005; Stewart and Stringer 2012). For instance, climatic changes such as those driven by ice ages can deeply modify the spatial distribution and genetic diversity of a species (Arenas, et al. 2012; Hewitt 2000). Even though AMH genetic diversity could have been impacted by cold periods (Arenas, et al. 2013; Hewitt 2000), the specific genetic effect of the last glacial age that occurred between 13-29 Kya ago (Straus 1991) has been little investigated. Similarly to what happened with many other species during the Last Glacial Maximum (LGM) (Lahr and Foley 1998), modern humans might have retreated to refuge areas such as southern Europe as well as southern and south-eastern Asia (Banks, et al. 2008; Stewart and Stringer 2012), and the extent of genetic differentiation between populations from different refugia might have increased during this period.

A consensual aspect of the migration process out of Africa is the relatively short time (~10-15 ky) it took for AMH to reach regions as remote as Southeast Asia and Oceania (see e.g., (Mellars 2006a) and references therein). It has been suggested that the early arrival of humans in Australia, which involved crossing water barriers, was only possible by the use of watercrafts such as boats or rafts (Balme 2013; Stringer 2000). Therefore, the migration of AMH to Oceania might have occurred through a form of long-distance dispersal (LDD). Archaeological evidence also suggests that artefacts associated with modern human behavior and hunting might have been spread over long distances, supporting likewise the existence of long-distance migration events during AMH history (McBrearty and Brooks 2000). Thus, although LDD has been associated to early modern behavior (Davidson and Noble 1992; Norton and Jin 2009), its role in the spread of Upper Paleolithic hunter-gatherer populations along the Eurasian (or American) coastlines is poorly understood. Similarly, its role in shaping human diversity has been little investigated as human colonization history has been mainly modeled under simple short-distance migration models (Fagundes, et al. 2007; Gutenkunst, et al. 2009; Hey 2005; Ray, et al. 2010b; Veeramah, et al. 2012). Whereas a few studies have used spatially-explicit models to describe the spread of humans (e.g. Eriksson, et al. 2012; Liu, et al. 2006;

Prugnolle, et al. 2005) and to take environmental heterogeneity into account (Ray, et al. 2005), LDD events were not considered. In this study, we aim at better understanding which demographic forces have affected human evolution and to more precisely assessing whether LDD and climatic changes have played a significant role in shaping current patterns of human genetic diversity. We analyze a large population dataset representative of Old World diversity under an Approximate Bayesian Computation (ABC) (Beaumont 2010; Beaumont, et al. 2002) framework. This procedure allows us to estimate the relative probabilities of several scenarios of human dispersal out of Africa, ranging from a very simple spatial expansion to more complex models including LDD events and a LGM-induced range contraction, or Neolithic-induced demographic expansions.

Results

We integrated spatially-explicit simulations into an Approximate Bayesian computation (ABC) approach (Beaumont 2010; Beaumont, et al. 2002) to assess if long-distance dispersal (LDD) and LGM-induced range contractions have shaped patterns of genetic diversity within and between human populations. Figure 1 shows key features of our tested spatially explicit scenarios, such as the initial Sub-Saharan range of modern humans (Fig. 1A), the spread out of Africa into Eurasia with (Fig. 1B) or without (Fig. 1A) LDD events, and the potential Eurasian refuge areas during the LGM (Fig. 1C).

We first tested four scenarios of modern human range expansion throughout the Old World. The simplest model (abbreviated as *noLDDnoRC*) consisted in a simple range expansion of modern humans from Eastern Africa (Prugnolle, et al. 2005), where migrations only occurred between neighboring populations, and for which there was no LDD and no range contraction. The second model (abbreviated *noLDDRC*) is identical to the first one, except that we modeled a range contraction during the LGM towards southern refugia in Europe and Asia (as shown in Fig. 1C). The third model (abbreviated *LDDnoRC*) is like the first model, except that LDD events occurred towards empty or already occupied demes. Finally, the fourth model (abbreviated *LDDRC*) is a model that included both LDD events and a range contraction (see Material and Methods for more details on model implementation). Even though these four evolutionary scenarios are a great simplification of the real demographic history of our species, they should nevertheless capture key aspects of human evolution. To evaluate the relative fit of each of these four models, we looked at the molecular diversity at 87 microsatellite markers passing strict quality filtering (see Material and Methods) across 22 populations selected to cover the Old World relatively evenly (as shown in Fig. 1D). The relative posterior probability of each model was assessed under an Approximate Bayesian Computation framework (Beaumont, et al. 2002(Beaumont 2008; Beaumont, et al. 2002). We thus fitted simulated microsatellite diversity at 50 loci with that observed at a thousand combinations of 50-loci resampled (bootstrapped) from the original set of 87 microsatellite markers (see Material and Methods). Note that the computed model posterior probabilities are only relative to tested models and do not integrate over all possible models of evolution of AMH.

Models without LDD are clearly rejected

Model posterior probability distributions

Using a multivariate logistic regression approach (Beaumont 2008), we estimated the posterior probabilities of the four models described above for each of the 1000 bootstrap datasets (see Material and Methods). We find that the two evolutionary scenarios without LDD have extremely low posterior probabilities (median Pr(*noLDDnoRC*) = $3.38\text{x}10^{-23}$; median Pr(*noLDDRC*) = $1.74\text{x}10^{-49}$, Fig. 2), and are thus clearly not supported by the data. Contrastingly, the *LDDRC* model, including both a LGM-induced range contraction and long distance dispersal is very strongly supported by the data with posterior probabilities very close to one (median Pr(*LDDRC*) = 0.9995; bootstrap CI95%: 0.9922-1.0000), and with the highest posterior probability among all 1000 bootstrap datasets (Fig. 2), whereas the *LDDnoRC* model, not including any LGM contraction, is much less supported (median Pr(*LDDnoRC*) = 0.0005, bootstrap CI95%: 3.23x10^{-05}-0.00789).

Accuracy of model choice

We checked that our model choice procedure could correctly recover the true model by simulating genetic datasets (referred to hereafter as Pseudo-Observed Data Sets or PODS) under each model and estimating their posterior probabilities in the same way as for the observed data sets. We find that we can almost perfectly identify models without LDD (99.5% of correct identifications, supplementary Table S1, supplementary Fig. S1, Supplementary Material). For PODS generated under models with LDD, correct model identification drops to 86.8% and 87.9% for the *LDDnoRC* and *LDDRC* models, respectively (supplementary Fig. S1, supplementary Table S1, Supplementary Material), but in those two cases the misidentified model is almost always (>99.5%) the alternative model with LDD. It implies that the presence or absence of the range contraction might be difficult to assess in presence of LDD. We also note that the proportion of correct assignment drops for higher levels of LDD (supplementary Fig. S2, Supplementary Material), suggesting that high frequencies of LDD can erase signals of range contractions. The fact that data sets generated under models without LDD are never assigned to scenarios with LDD (supplementary Table S1, supplementary Fig. S1, Supplementary Material) suggests that LDD creates genetic signatures that cannot be reproduced by models without LDD, such as more uniformly distributed allele frequencies (see below). In addition, our model validation analysis reveals that when the threshold for model assignment is large (>85%) model miss-assignment is almost negligible (supplementary Fig. S1, supplementary Table S1, Supplementary Material), even between models including LDD.

Goodness of fit

We checked whether observed patterns of diversity could be reproduced under our simulated scenarios. We thus directly compared the observed summary statistics (SSs) to those generated under the different models. Overall we find that most observed SSs can be reproduced under the four models (Fig. 3, supplementary Fig. S3A, Supplementary Material). A closer examination of the SS reveals that many of them are much better reproduced by models with LDD (SSs with red labels on Fig. 3). Indeed, models with LDD allow populations to show higher levels of heterozygosity more similar to the observed values, as well as lower levels of population differentiation, as measured by F_{ST}, within and between geographic regions outside Africa (Fig. 3 red labels, supplementary Fig. S3B, Supplementary Material). On the other hand, models with LDD tend to erase signatures of bottlenecks as they cannot reproduce a statistic indicative of bottleneck (NGW, Garza and Williamson 2001) in Europe (NGW_ENA blue labeled on Fig. 3), and they underestimate the amount of differentiation between Africa and Europe/Central Asia (FST_AFR-ENA and FST_AFR_CAS blue labeled on Fig. 3, see also supplementary Fig. S3C, Supplementary Material). As shown in Fig. 3, adding a LGM-induced range contraction in the model with LDD (LDDRC) leads to a better fit of the average number of alleles in Central Asian populations (K_CAS), of FST among Central Asian populations (FST_CAS), and (to a lesser extent) of FST among European populations (FST_ENA).Indeed, European and Central Asian populations tend to be overly differentiated in models without range contractions (see supplementary Fig. S3B, Supplementary Material). We also find that the inclusion of range contractions in models with and without LDD tend to generate a steeper decrease in the average number of alleles per locus as one moves to higher latitudes, but this pattern does not allow us to discriminate between models (see supplementary Fig. S4, Supplementary Material). Finally, all models tend to generate a decay of population heterozygosity with distance from Africa, even though this decay is steeper than, but still overlapping with, the observed data (slope and slope_levant in supplementary Fig. S3A, Supplementary Material). In summary, the genetic signature of models with LDD are elevated levels of diversity within populations and low levels of population differentiation that are very consistent with patterns observed in the empirical data.

LDD to empty habitats were disfavored during the colonization of Eurasia

During a range expansion, LDD events can either lead to the colonization of a new subpopulation (deme) if migrants are sent to empty demes or to gene flow between already occupied demes. In order to assess which type of LDD events are important during the colonization phase, we contrasted the two previously favored *LDDRC* and *LDDnoRC* models with scenarios where LDD could only occur towards empty places (**ep*) or only towards already occupied places (**op*) (Fig. 4) (see Material and Methods for details). Interestingly, we find that models where LDD events are only occurring towards empty places are strongly disfavored, and that the best supported model is one where LDD only occurs towards already occupied places (*LDDRCop*). This suggests that LDD events have been especially important to maintain genetic cohesion over long distances in already colonized regions of the world and that the colonization of Eurasia did not proceed by bursts of long distance dispersal. Moreover, the fact that a model with LDD events occurring only between occupied places is found vastly superior to *LDDRC* (where LDD can occur to both empty and occupied places), imply that LDD to empty habitats ahead of the wave front have been very rare during the colonization of Eurasia.

A model validation study shows that we have very high power (83.2%) to recognize the *LDDRCop* as the best model when it is true, and on average only ~3.3% chance of misidentifying it as the best model when one of the other five models is true (see supplementary Fig. S5, supplementary Table S2, Supplementary Material). This result reinforces the view that LDD to empty places either did not occur at all or occurred at much lower rates than LDD between already occupied habitats. In addition, we find that our ability to distinguish this model from a similar one involving no LGM-induced range contraction is increased (supplementary Fig. S5, supplementary Table S2, Supplementary Material) as compared to the case where LDD occur to any deme. These results strengthen support for the existence of refuge areas into which many individuals migrated during the last LGM.

Although all six scenarios with LDD produced very similar levels of genetic diversity within populations (supplementary Fig. S6A, Supplementary Material), irrespective of how LDD was implemented, *op scenarios display better fitting values of F_{ST} between non-African populations (supplementary Fig. S6B, Supplementary Material), and a more compatible extent of differentiation between African and European populations (F_{ST}_AFR-ENA, supplementary Fig. S6B, Supplementary Material). To access the goodness-of-fit of our LDD models and due to the high dimensionality of the SS space, we also examined the fit of the joint distribution of the simulated SSs to the empirical data by performing PCA analyses on the posterior distribution of the SSs generated under all models ("posterior predictive checks", supplementary Fig. S7, Supplementary Material). Besides the general good fit of all scenarios involving LDD, on average the *LDDRCop* matches most of the six PCs combinations (2D p-values ranging from 0.69 to 0.98) better than any of the remaining scenarios (*LDDnoRC* 2D p-values: 0.65-0.88; *LDDnoRCep* 2D p-values: 0.29-0.99; *LDDnoRCop* 2D p-values: 0.54-0.98; *LDDRC* 2D p-values: 0.53-0.85; *LDDRCep* 2D p-values: 0.17-0.99, supplementary Fig. S7, Supplementary Material).

Inferred demographic parameters

Most demographic parameters obtained under our best model (*LDDRCop*, Table 1, supplementary Fig. S8, Supplementary Material) are in very good agreement with previous estimates obtained under different approaches and data sets. An ancestral size of about 12,000 individuals is virtually identical to that obtained by an ABC analysis (Wegmann, et al. 2009) of tetranucleotide microsatellite diversity in Africa. A mean carrying capacity of about 1036 is very close to a previously estimated average size for hunter-gatherers of 839 (Hamilton, et al. 2007) and within a previously estimated range of 600-1200 (Deshpande, et al. 2009). Moreover, assuming a 1:3 ratio for effective:census population size (Deshpande, et al. 2009; Liu, et al. 2006) and a deme size of 150x150 km, the reported *K* can be translated into a density of 0.14 individuals per km^2, well within the 0.01-0.35 previously reported ranges for hunter-gatherer societies (Liu, et al. 2006, and references therein). Our migration rate estimate of 0.15 is also close to the colonization rate of 0.13 estimated from the decay of heterozygosity with distance from Africa (Deshpande, et al. 2009). Our estimate of migration rates, LDD proportions and effective sizes suggest that about eight long distance migrants (about 0.8 % of the local effective size) left each deme per generation and travelled an average of 717 km. The trinucleotide mutation rate of 1.72e-4 is slightly lower than recent direct estimates obtained from pedigrees (2.7e-4 for dinucleotides, (Sun, et al. 2012)), but in the range of previous estimates (0.9e-4-2.45e-4, (Wegmann and Excoffier 2010; Wegmann, et al. 2009)). The time of exit out of sub-Saharan Africa (posterior mean = 66 kya) is slightly older than previous estimates of 50-60,000 years, even though the lower limit of our confidence interval includes these more recent estimates (95%HPDI: 48-80 kya). The estimated growth rate (r) of 0.56 per generation (95%HPDI: 0.20-0.92) is also in very good agreement to previous estimates (Eriksson, et al. 2012) and, once more, well within the range estimated for hunter-gatherer populations ($0.3<r<0.9$, (Liu, et al. 2006) and references therein). Our estimate for the beginning of the colonization of Africa (mean = 95 kya) by modern humans are more recent than those obtained by Eriksson *et* al. (Eriksson, et al. 2012) and considerably more recent than the oldest remaining of AMH found in Africa (ref. 23,24 in (Liu, et al. 2006)), nevertheless, the upper limit of our confidence intervals (95%HPDI 80-121 kya) overlaps with previous estimates obtained from genetic data (Fagundes, et al. 2007; Voight, et al. 2005).

Discussion

Necessity and costs of LDD during human evolution

Long distance dispersal events have been often suspected in humans (Cavalli-Sforza, et al. 1994) and even predicted from patterns of diversity (e.g. Hellenthal, et al. 2014; Lawson, et al. 2012). However, their importance in human evolution has never been formally tested in a spatially-explicit context, and their global effect on patterns of human genetic diversity have not been examined in detail. We show here that models of human expansions in Eurasia without LDD events can be clearly rejected, and that LDD events allowed human populations to maintain large levels of genetic diversity despite their relatively low densities before the Neolithic (Table 1). Indeed, the occurrence of LD migration appears to prevent large divergence between continental groups (supplementary Fig. S3B, Supplementary Material), and it might therefore partly explain why rare and young haplotypes are more often shared between rather than within populations, as recently reported (Mathieson and McVean 2014). Even though it is difficult to pinpoint what exactly triggered LDD events in prehistoric times (e.g. curiosity, inbreeding avoidance, changes in available resources, seasonality, competition), it is worth noting that commercial trade could have promoted long distance migrations, as attested by a recent study (Smith, et al. 2015). Even though models including LDD are always more strongly supported than models without LDD, we find also strong evidence that no or very few migrants were sent ahead of the front during the colonization of Eurasia. Indeed, the best supported model is one where LDD would have only occurred between already colonized demes. This result might appear surprising, since higher dispersal abilities are supposed to evolve during range expansions (Phillips, et al. 2010; Shine, et al. 2011; Travis, et al. 2010), but several factors could have prevented or severely limited these types of LDD. First, isolated long distance migrants could have suffered from an Allee effect (Allee, et al. 1949; Travis and Dytham 2002) (which is not implemented in our simulations), and be quickly exposed to strong inbreeding depression and a reduction in fitness, which would have prevented the emergence of fast growing colonies ahead of the main expanding wave front. On one hand, this is compatible with the results of a recent study showing that expanding populations show a decrease in fitness due to the accumulation of deleterious mutations on the edge of the expansion (Peischl, et al. 2013). On the other hand, this is also in line with recent evidence showing that Allee effects can prevent loss of diversity in expanding populations (Roques, et al. 2012), by allowing migrants from the core to restore diversity lost by gene surfing (Edmonds, et al. 2004; Klopfstein, et al. 2006). Second, long distance migrants could have had problems in establishing successful colonies ahead of the modern human wave front if there had been strong competitions with archaic humans for accessing local resources. In this scenario, humans would have expanded only very progressively and interactions with archaic humans would have been limited to the edge of the expanding front (Currat and Excoffier 2011). Note that a strong competition with archaic humans would only slightly slow down the human expansion wave (see Table 1 in ref. (Currat, et al. 2008)), and low levels of archaic introgression, such as those found with Neanderthals in Eurasia (2-3% (Green, et al. 2010; Prufer, et al. 2014; Reich, et al. 2011)) would have very limited effects on the summary statistics used in this study. Overall, the fact that only certain types of LDD events might have been favored during (and after) human expansion suggests that LDD have evolved under opposite constraints. LDD within occupied areas would have restored diversity lost during range expansions, and allowed populations to maintain sufficiently high local genetic diversity despite low effective size, and an absence or very limited LDD events towards areas already occupied by other humans would have prevented direct competition with archaic populations. At this stage it is still unclear if sporadic LDD events have occurred ahead of the expanding front but would have not left any trace in the current Eurasians.

The current study has been performed by comparing the degree of genetic diversity in human populations genotyped for a relatively small number (87) of microsatellite markers. It would be interesting to reproduce our study and confirm our results by using genomic data sets, covering a much larger extent of human diversity. However, currently available large SNP data sets (e.g. Li, et al. 2008) suffer from unclear ascertainment bias that is impossible to reproduce in our simulations. Even though some efforts are put into the making of large collections of worldwide human DNA sequence, there is still no matching collection of unbiased genetic data that covers the Old World as evenly as this microsatellite database. Forthcoming whole-genome sequence data will certainly allow us to use linkage disequilibrium (LD) information, which has been shown to provide new insights on recent evolutionary events (Auton, et al. 2009). However, LD information was not used in the present study since we analyzed a set of independent and unlinked loci. In any case, several studies have shown that microsatellites are informative for the demographic history of human (Eriksson, et al. 2012; Ray, et al. 2005; Ray, et al. 2010b) and non-human (Becquet and Przeworski 2007) populations. Moreover, our power studies show that these sets of 50 microsatellites can very well distinguish between models (supplementary Figs. S1 and S5, Tables S1 and S2, Supplementary Material), which makes us confident that our results would not drastically change if larger genomic data sets would be used in matching populations.

A recent Neolithic population size increase cannot mimic the effect of LDD events

Since we have seen that the main effects of LDD events are to increase local diversity and reduce population divergence, one could wonder if a global recent increase in population size could not reproduce these patterns. Indeed the increase in local population sizes that occurred after the Neolithic must have also increased the number of migrants exchanged between neighboring demes, which has been shown to increase local diversity and decrease F_{ST} between populations (Excoffier 2004). To check if the implementation of a Neolithic population increase would make a model without LDD more compatible with the observations, we implemented a Neolithic-driven population growth in the *noLDDRC* model (see Material and Methods for details), and estimated its posterior probability relative to the *noLDDRC* and the *LDDRC* models. Interestingly, we find that a model with a Neolithic expansion but without LDD has the lowest support (supplementary Fig. S9, Supplementary Material) and does not seem to improve the fit to the observed summary statistics (supplementary Fig. S10, Supplementary Material). This shows that our support for the occurrence of LDD events during the colonization of Eurasia is not due to the potential confounding effect of a widespread population growth. It also suggests that the loss of genetic diversity (and increased population differentiation) seen in models without LDD occurs very early during a range expansion, and that a late Neolithic increase in population diversity cannot compensate this loss. We note that simulating a Neolithic expansion can sometimes improve the fit to the data, but only when range contractions and LDD events are also included in the model. However, this improvement is marginal and does not occur for all bootstrap data sets (data not shown). It is also worth noting that we did not simulate here the effect of secondary range expansion of Neolithic farmers (NF) into hunter-gatherer (HGs) territories. Previous simulations of this process have shown in the case of Europe that if NF and HG population had even small amount of interbreeding, the gene pool of the current population should be mainly of HG origin (Currat and Excoffier 2005; Currat, et al. 2008). At odds with this finding, the first ancient DNA studies suggested that there had been an almost complete replacement of hunter-gatherers (HGs) by Near East farmers (NEF) during the Neolithic in Europe (Bramanti, et al. 2009; Haak, et al. 2010; Malmstrom, et al. 2009). However, this view has been revised as recent studies now show that there is a larger genetic dissimilarity between Early Neolithic farmers (ENF) and modern Europeans than between Later Neolithic farmers (LNF) and modern Europeans (Bollongino, et al. 2013; Haak, et al. 2015a; Lazaridis, et al. 2014; Pinhasi, et al. 2012; Wilde, et al. 2014). Also the ENF influence is more visible in southern than in northern modern Europeans, which show stronger similarities with HGs (Haak, et al. 2015a; Skoglund, et al. 2012). These observations are compatible with a scenario where 1) the northern Mediterranean coast acted as a migration corridor for Neolithic farmers (Pinhasi, et al. 2012), and 2) Europe could have been recolonized from South to North establishing a gradient of genetic exchange between NF and HG populations (Skoglund, et al. 2012). Whereas a massive migration of NF from the Middle-East and an almost complete replacement of HG populations would mimic the effect of LDD by reducing the overall genetic differentiation between European populations, current evidence suggest that this replacement did not really occur or that it was only a partial replacement with strong heterogeneities across Europe (Haak, et al. 2015b; Lazaridis, et al. 2014; Pinhasi, et al. 2012; Skoglund, et al. 2012; Wilde, et al. 2014). The settlement of Europe has thus been more complex than previously anticipated, with even some contribution of a remote Western Asian population into Europe (Haak, et al. 2015b), potentially by a LDD event. However, it would appear very difficult to accurately model all the region-specific details of the settlement of Europe, a task that would be even more challenging in other regions were the amount of information on past demographic events is even lower. Our spatially explicit simulations with LDD, which do not introduce a strong discontinuity between HG and NF populations, appear nevertheless as a reasonably good approximation of the evolutionary history of Eurasia.

Note that we have not only tested the effect of larger deme carrying capacities, but also of larger deme physical dimensions on the model choice procedure, since LDD could be artificially favored by our model choice procedure if the dimension of the demes is too small in our simulations. To explicitly test this possibility, we performed simulations under the *noLDDRC* and *LDDRC* models with demes occupying areas of 250×250 km (instead of 150×150 km), and we adjusted their carrying capacity to keep the same density per square km. We then computed the posterior probabilities of these new models relative to those with the original deme dimensions. As expected, increasing deme dimensions increases the average posterior probability of the *noLDDRC* scenario (supplementary Fig. S11, Supplementary Material), but it also increases the posterior probability of the *LDDRC* scenario, which remains the overall best model. Whereas a larger deme dimension improves the posterior probability of all models, we note that a 250 km-wide deme size is much larger than any previous estimates of deme dimensions reported for hunter-gatherer populations (Ammerman and Cavalli-Sforza 1984; Anderson and Gillam 2000; Gronenborn 1999; Hewlett, et al. 1982), and it is 2.5 times larger than deme dimensions used in previous spatially explicit simulations studies (Eriksson, et al. 2012; Ray, et al. 2005). Since scenarios with LDD are better supported even with large deme sizes, our conclusion that LDD events are necessary to explain patterns of human diversity should remain valid.

Evidence for LGM induced range contractions in Eurasia

Our results suggest that the LGM only marginally affected patterns of human genetic diversity. This is in line with the observation that the effect of range contractions is only obvious when there is little gene flow between neighboring demes (Arenas, et al. 2012), and subsequent LDD events could have thus lessen the effects of past range contractions. Whereas our model choice procedure provides support for models involving both LDD events and range contractions, range contractions do not seem to have a very pronounced effect on the summary statistics when LDD events are implemented, with the exception of the number of alleles found in Central Asians, the average F_{ST} between Central Asian populations and to a less extent the average F_{ST} among Europeans populations (Fig. 3, SSs indicated with asterisks). We also looked at the fit between simulated and observed data by considering pairs of SSs rather than considering single SSs, *i.e.* we examined two-dimensional distributions of SSs. Using this approach, we find that among 364 2D SS distributions for which at least one model generates SS joint distributions similar to the empirical data, the *LDDRC* model performed better in 152 cases, as compared to 119 cases for the *LDDnoRC* model (see Material and Methods for a detailed description of these computations). The *LDDRC* scenario thus generates 2D SS modal values that are more often closer to the centroid of distributions obtained by bootstrapping the observed data, but this better fit is relatively marginal, suggesting that the overall better fit of the scenario involving range contractions emerges by considering all statistics collectively. We note however that, after restricting LDD events to occupied places, the relative support of scenarios with range contractions increases (compare Fig. 2 and 4), and the power to correctly identify such a scenario when it is true is still larger than 95% (using a 0.85 posterior probability threshold - supplementary Table S2, Supplementary Material). Simulated F_{ST} values within Europe, within central Asia and between Europe and Central Asia are also better fitting observations when range contractions are implemented. Nevertheless, the fact that there is a larger difference between statistics generated with and without range contractions in the absence of LDD (supplementary Fig. S3A/B, Supplementary Material), and that long distance migrations decrease our power to detect range contractions (supplementary Fig. S2, Supplementary Material), suggest that LDD events can partially erase the signature of previous range contractions. We also note that our implementation of range contraction is very simple and quite arbitrary, as there is still a lot of uncertainties on the exact place of refuge areas during the LGM and on the dynamics and timing of these contractions. Along these lines, although the spatially explicit range expansion models tested here are more realistic than those implemented in previous studies (but see Eriksson, et al. 2012, for another spatially explicit simulation framework), they are still very simplistic relative to the likely complex demographic history of humans (Henn, et al. 2012). Additional sources of realism could include temporal and regional differences in local population sizes and migration rates related to environmental heterogeneity (Mona, et al. 2014; Ray, et al. 2008), biased sex dispersal (Benguigui and Arenas 2014), a decoupling of migration and colonization rates (Deshpande, et al. 2009), or kin and mass migration (Fix 2004). However such complex models could become rapidly over determined and more difficult to explore due to the large number of extra parameters to examine.

Conclusions

Whereas we find that LDD ahead of the wave front were not very important during the settlement of Eurasia, this conclusion might not be valid for all continents. There is indeed some evidence that Oceania has been colonized >50 Kya by long distance migration events, probably by people using boats to reach Australia (Balme 2013; Stringer 2000). Dispersal by boat has been used more recently for the colonization of the Pacific Ocean, and there are indications that the fast colonization of the Americas was made possible by fast coastal migrations (Reich, et al. 2012; Wang, et al. 2007), potentially over long distances. However, the fact that regions in which there is evidence for long distance migrations during the colonization process were devoid of any archaic humans (Oceania, Polynesia, the Americas), is in line with the view that LDD events ahead of the colonization front might have been prevented or limited by interspecific competition. It suggests that intrinsic effects such as inbreeding depression might not have been the main factors preventing LDD events ahead of the expansion front. Therefore the inland colonization of Eurasia has been a quite progressive process, where serial founder effects (Henn, et al. 2012) and gene surfing (Excoffier, et al. 2009) might have led to a progressive loss of diversity with distance from Africa (e.g. Prugnolle, et al. 2005).

However, LDD events from the core to the front might have quickly restored diversity and reshuffled the genetic diversity of populations in Eurasia. These LDD events might also explain why the gene pool of many human populations shows signals currently interpreted as admixture events between isolated populations (Moorjani, et al. 2013; Patterson, et al. 2012; Pickrell, et al. 2014) that could just represent normal patterns having been built since the exit of modern humans from Africa.

Material and methods

Population samples and genetic marker selection

Patterns of human genetic variation were summarized by selecting 22 worldwide populations from a larger microsatellite dataset including 848 loci (Tishkoff, et al. 2009). The 22 populations were selected to have at least 15 sampled individuals (except Mongola and Uygur, with n=10) and to have a relatively even geographic distribution over the Old world (Fig. 1D). We considered here only trinucleotide microsatellites to prevent a putative ascertainment bias effect recently reported for tetranucleotide repeat loci (Eriksson and Manica 2011). In addition, we filtered out those loci showing properties that did not fit the stepwise-mutation model (SMM) used in our simulations (e.g. showing evidence of indels or incomplete repeated motifs). We also removed loci with >5% missing data as well as compound microsatellites (Pemberton, et al. 2009). This resulting set of 87 microsatellites is referred to hereafter as the "filtered dataset".

Demographic simulations

An extended version of SPLATCHE2 (Ray, et al. 2010a) was used to simulate the colonization of anatomically modern humans (AMH) across the Old World, implementing long-distance dispersal (LDD) as well as range contraction (RC) (Arenas, et al. 2012; Ray and Excoffier 2010). SPLATCHE is a two-step simulator, in which the first-step concerns the simulation (forward in time) of the population demography, i.e. deme growth and migration history, and the second step consists in generating genetic diversity via coalescent simulations using the information recorded during the demographic step. Therefore, genetic drift occurs through the process of random sampling within the coalescent simulations (Excoffier, et al. 2009; Ray, et al. 2010a). We generated land surface by using the Hammer-Aitoff projection, which divides the Old World into 9,047 squared demes (subpopulations) of 22,500 km^2 each. Land bridges connecting the continent to some islands, such as Great Britain and Japan, have been artificially added to allow the settlement of these regions in scenarios without LDD.

We assumed that the range expansion started some $T_{STARTEXP}$ generations ago from a single deme (located in Addis Ababa, Ethiopia, following (Pritchard, et al. 2000) assumed to have an ancestral population size equal to Ne_{ANC} diploid individuals. Each generation, deme density is logistically regulated, with an intrinsic growth rate (r) and a carrying capacity (K), followed by a migration phase, in which $2N_tm$ migrants are sent to the four neighboring demes, where N_t stands for the haploid number of individuals within the deme at generation t. All model parameters are drawn from prior distributions shown in supplementary Table S3, Supplementary Material.

Long-distance dispersal

In our simulation framework, LDD events represent migration events to non-adjacent demes. The total number of long distance (LD) migrant genes is drawn from a binomial distribution $b(2N_t, m, LDD_{PROP})$, where LDD_{PROP} is the proportion of the total number of migrant genes ($2N_tm$) that are long distance dispersers. Once the number of LD migrant genes is computed each LD migrant moves X demes away, where X is sampled from a gamma distribution $Gamma(\alpha, \alpha/\bar{\mu})$ where α is the shape parameter and $\bar{\mu}$ is the average distance travelled by the migrant individuals. Both α and $\bar{\mu}$ are hyper-parameters drawn from uniform priors (see below), which ranges produces distributions ranging from exponential to almost normal. Therefore, our method is able to cover alternative LDD kernels. In order to avoid unrealistically long LD migration events, we arbitrarily set the maximum dispersal distance to 6 demes (900 km). See Ray and Excoffier (Ray and Excoffier 2010) for a more detailed description of the LDD distribution kernel.

Last Glacial Maximum (LGM) range contraction

The range contraction underlying the LGM was modeled by shrinking progressively the range of habitable areas for human populations. To model a LGM-induced range contraction, the onset of the range contraction (T_{SCONTR}) was set to 1,000 generations ago or 25,000 BP (assuming a 25y generation time), and its dynamics was modeled by simulating 30 range contraction events, two or four generations apart, in which two or one latitudinal layer of demes become uninhabitable by setting their carrying capacity to zero (see ref (Arenas, et al. 2012) for more details on this process). The number of generations between each contraction event and the width of each contraction (in number of demes) were chosen to fit the potential duration of the contraction process (120 generations or 3,000 y) and the potential geographical location of the refugia (see below). Populations were then restricted to refugia shown in Fig. 1C for 160 generations (4,000 y), and then free to migrate again to recolonize the Eurasian landmass. The refuge areas suitable for humans during the LGM were derived from paleo-vegetation maps of the LGM (Ray and Adams 2001). They consisted in: i) Southern Europe, represented as a continuum from the Iberian Peninsula to the Turkish coast; ii) a very narrow strip over the northern most part of Africa; iii) India and South Eastern China, which were also modeled as a continuous region; and iv) sub-Saharan Africa.

Coalescent simulations

During the forward demographic simulations, we recorded at each generation the deme densities, as well as the number of LDD and non-LDD immigrants. This demographic database was then used to perform coalescent simulations backward in time and to generate genetic data. Due to the computational cost of our spatially-explicit simulations, we simulated multi-locus genotypes for only 50 microsatellite loci, matching the number of individuals per sampled population (Fig. 1D), assuming a strict stepwise mutation model (SMM) and a mutation rate at all loci drawn from a uniform prior $STR_{MUTRATE} \sim U[5x10^{-5}\text{-}3x10^{-4}]$ (supplementary Table S3, Supplementary Material).

Main tested evolutionary scenarios

We first examined four alternative scenarios including LDD and range contraction or not. The simplest model examined in this study involves a single range expansion with no LDD and no range contraction (hereafter referred to as *noLDDnoRC*). It posits that modern humans emerged in Eastern Africa 3,200 to 6,000 generations BP (80-160 Kya). They then spread across Sub-Saharan Africa, where they remained until 1,200-3,000 generations (30-75Ky) before being allowed to migrate along the Nile river valley and the Red Sea coast to colonize the Eurasian continent (the Out-of-Africa event). Then, individuals could migrate over the Old World except the Arabic desert and the Himalayas, where carrying capacities were set to zero. We implemented two additional models derived from the *noLDDnoRC* model, by adding either a LGM-induced range contraction (the *noLDDRC* model) or LDD events (the *LDDnoRC* model), as described above. The fourth tested model includes both a range contraction and LDD events and is abbreviated as *LDDRC.* Note that here LDD migrants can either be sent to an empty or to an occupied deme, the former case being a founding event. All demographic parameters such as carrying capacities (K), growth rate (r), migration rate (m), LDD proportion (LDD_{PROP}), and all the parameters related to the movement of LDD migrants are drawn from prior distributions shown in supplementary Table S3, Supplementary Material. Note that all parameters related to the range contraction were fixed.

Additionally tested scenarios

In order to see if the Neolithic and its associated population size increase could generate patterns similar to those produced by LDD, we extended the *noLDDRC* to include a sudden population size increase 320 generations (8,000 y) ago. We thus included an additional parameter (*KNeo*~U[2500-7500], supplementary Table S3, Supplementary Material) representing the carrying capacity of post-Neolithic populations.

We also studied two alternative scenarios with LDD, where i) LDD could only occur towards empty places (demes), implying they would always correspond to a colonization event and ii) LDD could only occur towards already colonized demes. We use the **ep* (empty places) and **op* (occupied places) suffixes to refer to these two scenarios.

Approximate Bayesian Computations

We used approximate Bayesian computation (ABC) (Beaumont 2010; Beaumont, et al. 2002) to compute the relative posterior probabilities of the tested models via simulations. Genetic datasets matching the sample composition and the number loci used in this study were generated by: 1) drawing parameter values from prior distributions (supplementary Table S3, Supplementary Material); 2) performing demographic and coalescent simulations using SPLATCHE 2 (Ray, et al. 2010a); 3) recording summary statistics (SS) computed on simulated genetic data using arlsumstat (Excoffier and Lischer 2010). All these steps were performed 100,000 times for each of the four main models using ABCtoolbox (Wegmann, et al. 2010). Due to the large computational costs of our simulations, other model comparisons were performed using fewer simulations. Summary statistics and model choice procedure are explained below.

Summary statistics (SS)

Simulated and observed genotype data were summarized by computing for each population the mean and the standard deviation (SD) over loci of the following statistics: number of alleles (K), heterozygosity (H) and modified Garza-Williamson statistics (NGW) (Fuller, et al. 2011). The SD of these summary statistics was also computed after pooling all samples (supplementary Table S4, Supplementary Material). Additionally, we computed F_{ST} between all pairs of populations taking into account the difference in number of repeat number between microsatellite alleles (R_{ST}). In order to reduce the number of SS, we averaged the SSs within four geographically-defined groups of populations (Fig. 1D, main text). Pairwise R_{ST} were also averaged within and between population groups. We measured the decay of population heterozygosity with least-cost geographic distance from the origin of the expansion by computing the slope of the fitted linear regression. Since with LDD, migrant genes can cross short sea tracts (e.g. migrate directly from the Horn of Africa to the Arabian Peninsula, or from Morocco into the Iberian Peninsula), we computed the slope of the regression on two geographic distance matrices. In the first matrix, geographic distances between populations were computed assuming that migrants could only migrate out of Africa through the Levant, and in the second matrix distances were computed by allowing migration out of Africa also through the Horn of Africa and Gibraltar. The 39 SSs used to perform ABC model choice are listed in supplementary Table S4, Supplementary Material.

Model choice

To perform model choice we first computed the Euclidian distance between the observed and simulated SS under the different models, and then retained the 2-10% (tolerance) of the simulations closest to the observed dataset, depending on the total amount of simulations performed under different models. We then computed model posterior probabilities using the multivariate weighted logistic regression (MLR), introduced by Beaumont (Beaumont 2008; Fagundes, et al. 2007), where the model (M) indicator (1,..,n_M) is considered a categorical variable Y and logistic regression coefficients (β_j^T) for the model j are estimated in $P(Y = j \mid S) = \frac{e^{\left(\beta_j^T S\right)}}{\sum_{i=1}^{M} e^{\left(\beta_i^T S\right)}}$, where S are the observed summary statistics. The function vglm() and predictvgml() from the R package VGAM were used to compute the fitted coefficients and model posterior probabilities, respectively (see the following references (Beaumont 2008; Fagundes, et al. 2007) for more details). The MLR was performed on 1,000 observed datasets obtained by resampling 50 microsatellites without replacement from the "filtered dataset", providing confidence intervals for these posterior probabilities representing the influence of the choice of 50 microsatellites on our results. The MLR was computed with the R function *Calmod*, developed by M. Beaumont and available on https://code.google.com/p/popabc/.

Parameter inference

Parameters of the *LDDRCop* model were estimated via a conventional ABC procedure (Leuenberger and Wegmann 2010; Wegmann, et al. 2010). Partial least squares (PLS) components were first computed on Box-Cox transformed SSs to decrease the dimensionality of the SS space and to maximize the linearity between SS and parameters. The retained number of PLS components (12) was chosen such that the addition of more PLS would not decrease the root mean squared error (RMSE, see (Ray, et al. 2010b) for more details and supplementary Fig. S8, Supplementary Material). Parameter posterior distributions were computed from the 2,000 simulations (out of 50,000 simulations) closest to the observed data and under a general linear model (GLM) approximating the likelihood function, as implemented in ABCtoolbox (Wegmann, et al. 2010).

Power of the model choice procedure

To test if we have enough power to distinguish between alternative models, we used 1,000 randomly chosen simulations, referred to hereafter as pseudo-observed data sets (PODS), and computed model posterior probabilities for each PODS, as described above. A PODS was assigned to one of the models if its relative probability was larger than either 50% or 85%, and considered unassigned otherwise. This model validation procedure was carried out for the four main alternative models (supplementary Fig. S1, Supplementary Material) and for the six scenarios with different types of LDD events (supplementary Fig. S5, Supplementary Material).

Goodness of fit

We investigated the fit between the empirical (bootstrap replicates) and the simulated summary statistics generated under all models tested in this study. The fit was investigated by looking at individual SSs (Fig. 3, supplementary Fig. S3A/B/C, supplementary Fig. S6A/B, supplementary Fig. S10, supplementary Fig. S12, Supplementary Material), and/or at the 2D joint distributions of all pairwise combinations of the different SSs (data not shown). When investigating 2D joint distributions, we computed the 2D densities across the 1000 observed bootstrap replicates and then computed the p-values of the 2D highest density points (mode) obtained from the 2D joint distribution of the SSs recovered by retaining the best 2% simulation generated under alternative models. The 2D densities and p-values were computed following Daub et al. (Daub, et al. 2014). R scripts to perform these analyses are available upon request.

We also performed posterior predictive checks to verify that the selected scenarios can reproduce the empirical data. We first performed parameter inference using the same procedure as for the *LDDRCop* model, described above. We then sampled 1,000 parameter vectors from the obtained posterior distributions, and used them as input parameters for SPLATCHE2 (Ray, et al. 2010a) to generate 1,000 new genetic datasets for each demographic scenario. These "posterior datasets" were then summarized via Principal Component Analysis (PCA) in the case of the six models with LDD. The first four PCs were plotted against each other, and for each 2D plot we projected the observed dataset and computed the curves for the 95%, 75% and 50% envelopes of the posterior 2D predictive distributions (supplementary Fig. S7, Supplementary Material). The 2D densities and p-values were computed as described above.

Acknowledgments

We thank two anonymous reviewers for their constructive comments that improved the quality of the manuscript. This work was supported by the Swiss National Science Foundation (1003A-143393 to L.E), the EU/FP7 (226852 (SCALES) to LE) and the Portuguese Foundation for Science and Technology (SFRH/BD/73150/2010 to I.A).

References

Allee WC, Emerson AE, Park O, Schmidt KP. 1949. Principles of Animal Ecology. Philadelphia, PA, USA.

Ammerman AJ, Cavalli-Sforza LL. 1984. The Neolithic transition and the genetics of populations in Europe: Princeton University Press.

Anderson DG, Gillam JC 2000. Paleoindian colonization of the Americas: implications from an examination of physiography, demography, and artifact distribution. American Antiquity 65: 43-66.

Arenas M, Ray N, Currat M, Excoffier L 2012. Consequences of range contractions and range shifts on molecular diversity. Molecular biology and evolution 29: 207-218.

Arenas M, Francois O, Currat M, Ray N, Excoffier L 2013. Influence of admixture and paleolithic range contractions on current European diversity gradients. Molecular biology and evolution 30: 57-61.

Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, Indap A, Wright MH, Degenhardt JD, Gutenkunst RN 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res 19: 795-803.

Balme J 2013. Of boats and string: The maritime colonisation of Australia. Quaternary International 285: 68-75.

Banks WE, d'Errico F, Peterson AT, Vanhaeren M, Kageyama M, Sepulchre P, Ramstein G, Jost A, Lunt D 2008. Human ecological niches and territories during the LGM in Europe derived from an application of eco-cultural niche modeling. Journal of Archaeological Science 35: 481-491.

Beaumont M 2008. Joint determination of topology, divergence time, and immigration in population trees. In: Matsumura S, Renfrew PFC, editors. Simulation, genetics and human prehistory. Cambridge:McDonald Institute for Archeological Research: 134–154.

Beaumont MA, Zhang W, Balding DJ 2002. Approximate Bayesian Computation in Population Genetics. Genetics 162: 2025-2035.

Beaumont MA 2010. Approximate Bayesian Computation in Evolution and Ecology. Annual Review of Ecology, Evolution, and Systematics 41: 379-406.

Becquet C, Przeworski M 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Res 17: 1505-1519.

Benguigui M, Arenas M 2014. Spatial and temporal simulation of human evolution. Methods, frameworks and applications. Curr Genomics 15: 245-255.

Bollongino R, Nehlich O, Richards MP, Orschiedt J, Thomas MG, Sell C, Fajkošová Z, Powell A, Burger J 2013. 2000 Years of Parallel Societies in Stone Age Central Europe. Science 342: 479.

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, Antanaitis-Jacobs I, Haidle MN, Jankauskas R, Kind C-J, Leuth F, Terberger T, Hiller J, Matsumura S, Forster P, Burger J 2009. Genetic discontinuity between local Hunter-gatherers and Central Europe's first farmers. Science 326: 137-140.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The History and Geography of Human Genes. Princeton, NJ: Princeton University Press.

Currat M, Excoffier L 2005. The effect of the Neolithic expansion on European molecular diversity. Proceedings. Biological sciences / The Royal Society 272: 679-688.

Currat M, Ruedi M, Petit RJ, Excoffier L 2008. The hidden side of invasions: massive introgression by local genes. Evolution; international journal of organic evolution 62: 1908-1920.

Currat M, Excoffier L 2011. Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. Proc Natl Acad Sci U S A 108: 15129-15134.

Daub JT, Dupanloup I, Robinson-Rechavi M, Excoffier L 2014. Inference of evolutionary forces acting on human biological pathways. BioRxiv:10.1101/009928.

Davidson I, Noble W 1992. Why the First Colonisation of the Australian Region Is the Earliest Evidence of ModernHuman Behaviour. Archaeology in Oceania 27: 135-142.

Deshpande O, Batzoglou S, Feldman MW, Cavalli-Sforza LL 2009. A serial founder effect model for human settlement out of Africa. Proc Biol Sci 276: 291-300.

Edmonds CA, Lillie AS, Cavalli-Sforza LL 2004. Mutations arising in the wave front of an expanding population. PNAS 101: 975-979.

Eriksson A, Betti L, Friend AD, Lycett SJ, Singarayer JS, von Cramon-Taubadel N, Valdes PJ, Balloux F, Manica A 2012. Late Pleistocene climate change and the global expansion of anatomically modern humans. Proc Natl Acad Sci U S A 109: 16089-16094.

Eriksson A, Manica A 2011. Detecting and Removing Ascertainment Bias in Microsatellites from the HGDP-CEPH Panel. Genes Genomes Genetics 1: 479-488.

Excoffier L 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. Molecular ecology 13: 853-864.

Excoffier L, Foll M, Petit RJ 2009. Genetic Consequences of Range Expansions. Annual Review of Ecology, Evolution, and Systematics 40: 481-501.

Excoffier L, Lischer HE 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular ecology resources 10: 564-567.

Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L 2007. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci U S A 104: 17614-17619.

Fix AG 2004. Kin-structured migration: causes and consequences. Am J Hum Biol 16: 387-394.

Fuller DQ, Boivin N, Hoogervorst T, Allaby R 2011. Across the Indian Ocean: the prehistoric movement of plants and animals. Antiquity 85: 544-558.

Garza JC, Williamson EG 2001. Detection of reduction in population size using data from microsatellite loci. Molecular Ecology 10: 305–318.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, al. e 2010. A draft sequence of the Neandertal genome. Science 328: 710-722.

Gronenborn D 1999. A variation on a basic theme: the transition to farming in southern central Europe. Journal of world prehistory 13: 123-210.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5: e1000695.

Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, Der Sarkissian CS, Brandt G, Schwarz C, Nicklisch N, Dresely V, Fritsch B, Balanovska E, Villems R, Meller H, Alt KW, Cooper A, Members of the Genographic C 2010. Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PLoS Biol 8: e1000536.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, al. e 2015a. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. Nature advance online publication. doi: 10.1038/nature14317

Hamilton MJ, Milne BT, Walker RS, Burger O, Brown JH 2007. The complex structure of hunter-gatherer social networks. Proc Biol Sci 274: 2195-2202.

Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, Myers S 2014. A genetic atlas of human admixture history. Science 343: 747-751.

Henn BM, Cavalli-Sforza LL, Feldman MW 2012. The great human expansion. Proc Natl Acad Sci U S A 109: 17758-17764.

Hewitt G 2000. The genetic legacy of the Quaternary ice ages. Nature 405: 907-913.

Hewlett B, van de Koppel JM, Cavalli-Sforza LL 1982. Exploration ranges of Aka pygmies of the Central African Republic. Man 17: 418-430.

Hey J 2005. On the Number of New World Founders: A Population Genetic Portrait of the Peopling of the Americas. PLoS Biol 3: e193.

Klopfstein S, Currat M, Excoffier L 2006. The fate of mutations surfing on the wave of a range expansion. Mol Biol Evol 23: 482-490.

Lahr MM, Foley RA 1998. Towards a Theory of Modern Human Origins: Geography, Demography, and Diversity in Recent Human Evolution. YEARBOOK OF PHYSICAL ANTHROPOLOGY 41: 137–176.

Lawson DJ, Hellenthal G, Myers S, Falush D 2012. Inference of population structure using dense haplotype data. PLoS Genet 8: e1002453.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, al. e 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513: 409-413.

Leuenberger C, Wegmann D 2010. Bayesian computation and model selection without likelihoods. Genetics 184: 243-252.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-1104.

Liu H, Prugnolle F, Manica A, Balloux F 2006. A geographically explicit genetic model of worldwide human-settlement history. Am J Hum Genet 79: 230-237.

Malmstrom H, Gilbert MT, Thomas MG, Brandstrom M, Stora J, Molnar P, Andersen PK, Bendixen C, Holmlund G, Gotherstrom A, Willerslev E 2009. Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. Curr Biol 19: 1758-1762.

Mathieson I, McVean G 2014. Demography and the Age of Rare Variants. PLoS Genet 10.

McBrearty S, Brooks AS 2000. The revolution that wasn't: a new interpretation of the origin of modern human behavior. Journal of human evolution 39: 453-563.

Mellars P 2006a. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. Science 313: 796-800.

Mellars P 2006b. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. Proc Natl Acad Sci U S A 103: 9381-9386.

Mona S, Ray N, Arenas M, Excoffier L 2014. Genetic consequences of habitat fragmentation during a range expansion. Heredity 112: 291-299.

Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, Berger B, Reich D, Singh L 2013. Genetic evidence for recent population mixture in India. Am J Hum Genet 93: 422-438.

Norton CJ, Jin JJH 2009. The evolution of modern human behavior in East Asia: Current perspectives. Evolutionary Anthropology: Issues, News, and Reviews 18: 247-260.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D 2012. Ancient Admixture in Human History. Genetics 192: 1065–1093

Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L 2013. On the accumulation of deleterious mutations during range expansions. Molecular ecology 22: 5972-5982.

Pemberton TJ, Sandefur CI, Jakobsson M, Rosenberg NA 2009. Sequence determinants of human microsatellite variability. BMC genomics 10: 612.

Phillips BL, Brown GP, Shine R 2010. Life-history evolution in range-shifting populations. Ecology 91: 1617-1627.

Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A 111: 2632–2637.

Pinhasi R, Thomas MG, Hofreiter M, Currat M, Burger J 2012. The genetic history of Europeans. Trends Genet 28: 496-505.

Pritchard JK, Stephens M, Donnelly P 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics 155: 945–959.

Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, al. e 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-49.

Prugnolle F, Manica A, Balloux F 2005. Geography predicts neutral genetic diversity of human populations. Curr Biol 15: R159-160.

Ray N, Adams JM 2001. A GIS-based Vegetation Map of the World at the Last Glacial Maximum (25,000-15,000 BP). Internet Archaeology 11.

Ray N, Currat M, Berthier P, Excoffier L 2005. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. Genome Res 15: 1161-1167.

Ray N, Currat M, Excoffier L. 2008. Incorporating environmental heterogeneity in spatially-explicit simulations of human genetic diversity. In: Matsumura S, Forster P, Renfrew C, editors. Simulations, Genetics and Human Prehistory - A Focus on Islands. Cambridge: McDonald Institute for Archaeological Research. p. 103-117.

Ray N, Currat M, Foll M, Excoffier L 2010a. SPLATCHE2: a spatially-explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics 26: 2993-2994.

Ray N, Excoffier L 2010. A first step towards inferring levels of long-distance dispersal during past expansions. Molecular ecology resources 10: 902-914.

Ray N, Wegmann D, Fagundes NJ, Wang S, Ruiz-Linares A, Excoffier L 2010b. A statistical evaluation of models for the initial settlement of the american continent emphasizes the importance of gene flow with Asia. Molecular biology and evolution 27: 337-345.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, al. e 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053-1060.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, Garcia LF, Triana O, Blair S, Maestre A, Dib JC, Bravi CM, Bailliet G, Corach D, Hunemeier T, Bortolini MC, Salzano FM, Petzl-Erler ML, Acuna-Alonzo V, Aguilar-Salinas C, Canizales-Quinteros S, Tusie-Luna T, Riba L, Rodriguez-Cruz M, Lopez-Alarcon M, Coral-Vazquez R, Canto-Cetina T, Silva-Zolezzi I, Fernandez-Lopez JC, Contreras AV, Jimenez-Sanchez G, Gomez-Vazquez MJ, Molina J, Carracedo A, Salas A, Gallo C, Poletti G, Witonsky DB, Alkorta-Aranburu G, Sukernik RI, Osipova L, Fedorova SA, Vasquez R, Villena M, Moreau C, Barrantes R, Pauls D, Excoffier L, Bedoya G, Rothhammer F, Dugoujon JM, Larrouy G, Klitz W, Labuda D, Kidd J, Kidd K, Di Rienzo A, Freimer NB, Price AL, Ruiz-Linares A 2012. Reconstructing Native American population history. Nature 488: 370-374.

Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME, al. e 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet 89: 516-528.

Roques L, Garnier J, Hamel F, Klein EK 2012. Allee effect promotes diversity in traveling waves of colonization. Proc Natl Acad Sci U S A 109: 8828-8833.

Shine R, Brown GP, Phillips BL 2011. An evolutionary process that assembles phenotypes through space rather than through time. Proc Natl Acad Sci U S A 108: 5708-5711.

Skoglund P, Malmstrom H, Raghavan M, Stora J, Hall P, Willerslev E, Gilbert MT, Gotherstrom A, Jakobsson M 2012. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336: 466-469.

Smith O, Momber G, Bates R, Garwood P, Fitch S, Pallen M, Gaffney V, Allaby RG 2015. Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago. Science 347: 998-1001.

Stewart JR, Stringer CB 2012. Human evolution out of Africa: the role of refugia and climate change. Science 335: 1317-1321.

Straus LG 1991. Southwestern Europe at the Last Glacial Maximum. Current Anthropology 32: 189-199.

Stringer C 2000. Coasting out of Africa. Nature 405: 24-27.

Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, Stefansson K 2012. A direct characterization of human mutation based on microsatellites. Nat Genet 44: 1161-1165.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, al. e 2009. The Genetic Structure and History of Africans and African Americans. Science 324: 1035-1044.

Travis JM, Munkemuller T, Burton OJ 2010. Mutation surfing and the evolution of dispersal during range expansions. Journal of evolutionary biology 23: 2656-2667.

Travis JMJ, Dytham C 2002. Dispersal evolution during invasions. Evolutionary Ecology Research 4: 1119-1129.

Veeramah KR, Wegmann D, Woerner A, Mendez FL, Watkins JC, Destro-Bisol G, Soodyall H, Louie L, Hammer MF 2012. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Molecular biology and evolution 29: 617-630.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A 102: 18508-18513.

Wang S, Lewis C, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina J, Gallo C, Mazzotti G, Poletti G, Hill K, Hurtado A, Labuda D, Klitz W, Barrantes R, Bortolini

MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Llop E, Rothhammer F, Excoffier L, Feldman MW, Rosenberg NA, Ruiz-Linares A 2007. Genetic Variation and Population Structure in Native Americans. PLoS Genet 3.

Wegmann D, Excoffier L 2010. Bayesian inference of the demographic history of chimpanzees. Mol Biol Evol 27: 1425-1435.

Wegmann D, Leuenberger C, Excoffier L 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182: 1207-1218.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC bioinformatics 11: 116.

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Hollfelder N, Potekhina ID, Schier W, Thomas MG, Burger J 2014. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A 111: 4832–4837.

Tables

Table 1. Demographic parameters estimated under the best fitting model (*LDDRCop*).

Parameters	Mode	Mean	Median	95% HPDI[a]
Start of the initial expansion in Africa ($T_{STARTEXP}$)[b]	80,704	94,903	91,777	80,000-120,916
Out of sub-Saharan Africa expansion time (T_{OOA})[b]	73,568	65,924	67,477	48,276-80,000
Ancestral size (Ne_{ANC})[c]	10,327	11,795	11,386	5,000-19,098
Carrying capacity (K)[c]	826	1,036	992	50-1,992
LDD proportion (LDD_{PROP})	0.044	0.038	0.040	0.021-0.050
Growth rate (r)	0.429	0.561	0.545	0.200-0.919
Average number of demes travelled by LDD migrants (μ)	5.357	4.780	4.946	3.074-6.000
Gamma shape parameter – LDD distance (α)	1.209	1.251	1.249	0.567-1.943
Migration rate (m)	0.110	0.155	0.148	0.050-0.268
Number of migrants (Nm)[c]	3	93	76	3-241
Number of LDD migrants ($LDDNm$)[c]	8	8	8	0-15
Mutation rate ($\mathrm{STR}_{\mathrm{MUTRATE}}$)	1.74E-04	1.72E-04	1.72E-04	1.07E-04-2.36E-04

[a] HPDI – 95% highest posterior density interval.

[b] Times are in years, assuming a generation time of 25 years.

[c] Parameters estimates are reported in number of diploid individuals.

Figures

Figure legends

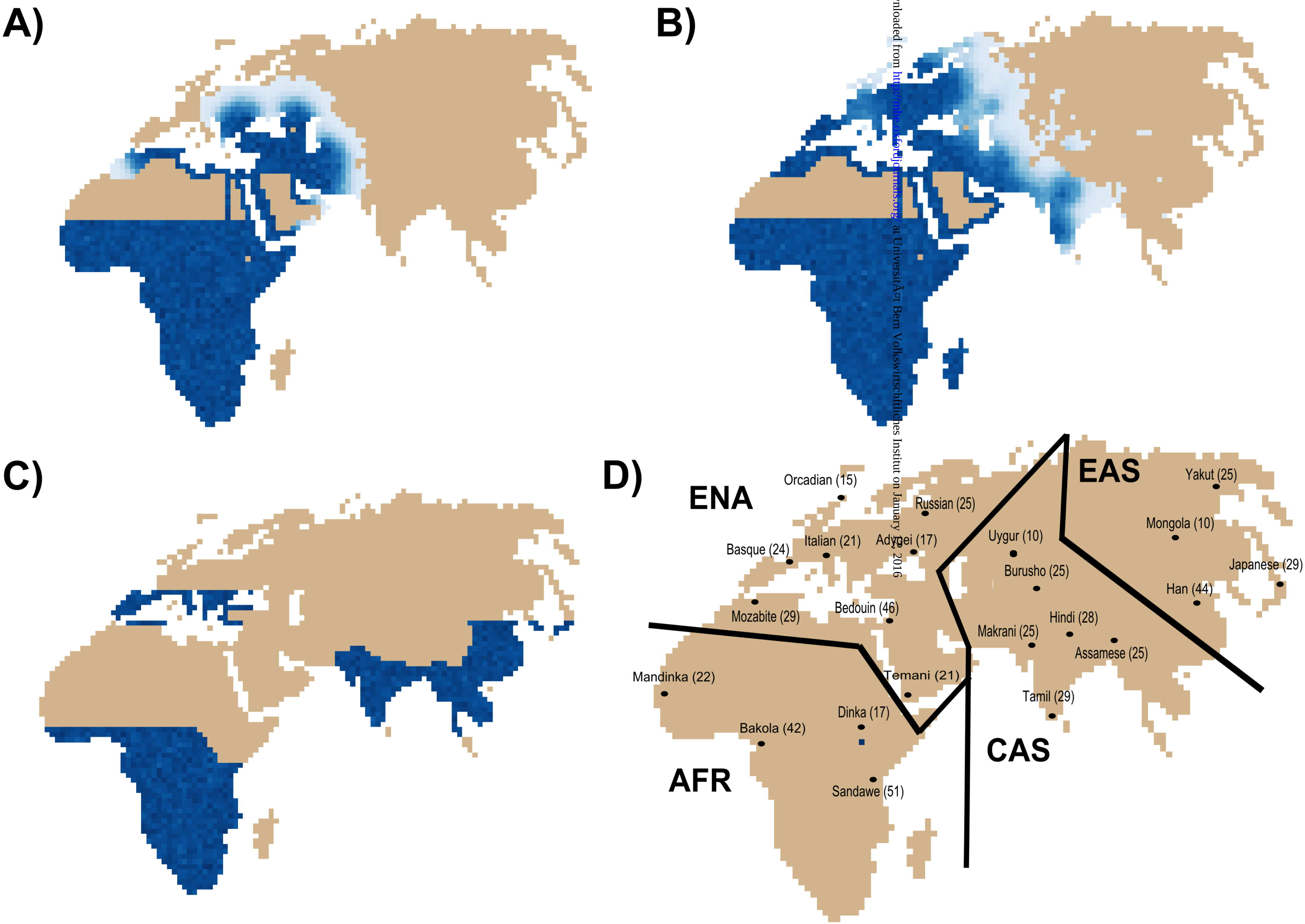
Figure 1. Main features of our spatially-explicit simulations, and sample locations. A) Exit out-Africa through the Nile valley and along North African coastlines. B) Colonization of Eurasia with visible LDD events ahead of the main wave front. C) LGM refuge areas in Europe and Asia and contraction of sub-Saharan Africa due to the extension of the Sahara. D) Locations and sizes of the 22 samples used in this study. The samples are found in four geographic regions abbreviated as AFR for sub-Saharan Africa, ENA for Europe, Near-East and North-Africa, CAS for Central Asia, and EAS for East Asia. The different shades of blue are proportional to population densities.

Figure 2. Distributions of the posterior probabilities of the four main scenarios of human expansions (*noLDDnoRC, noLDDRC, LDDnoRC, LDDRC*) obtained over the 1,000 bootstrap datasets. Model posterior probabilities were computed using the multivariate logistic regression (Beaumont 2008) on the 2% best simulations (closest to the empirical data) among 100,000 simulations per evolutionary scenario.

Figure 3. Fit between simulated and observed summary statistics for models *noLDDnoRC*, *noLDDRC, LDDnoRC* and *LDDRC*. We report here, for each SS, the mode of the 2% simulations closest to the observations retained for posterior probability estimations. SSs were standardized using the mean and the standard deviation of each SS obtained across the 1,000 bootstrap datasets. Thus, SS modal values closer to zero (grey dashed line) indicate better fit to the observed SSs. The 39 SSs are fully described in Table S4. We have highlighted in red SSs where LDD scenarios were better supported by the data, and in blue SSs where no-LDD scenarios were better supported. SSs better supported by the *LDDRC* than by the *LDDnoRC* model are indicated by asterisks. All SS distributions are shown in supplementary Fig. S3A, Supplementary Material.

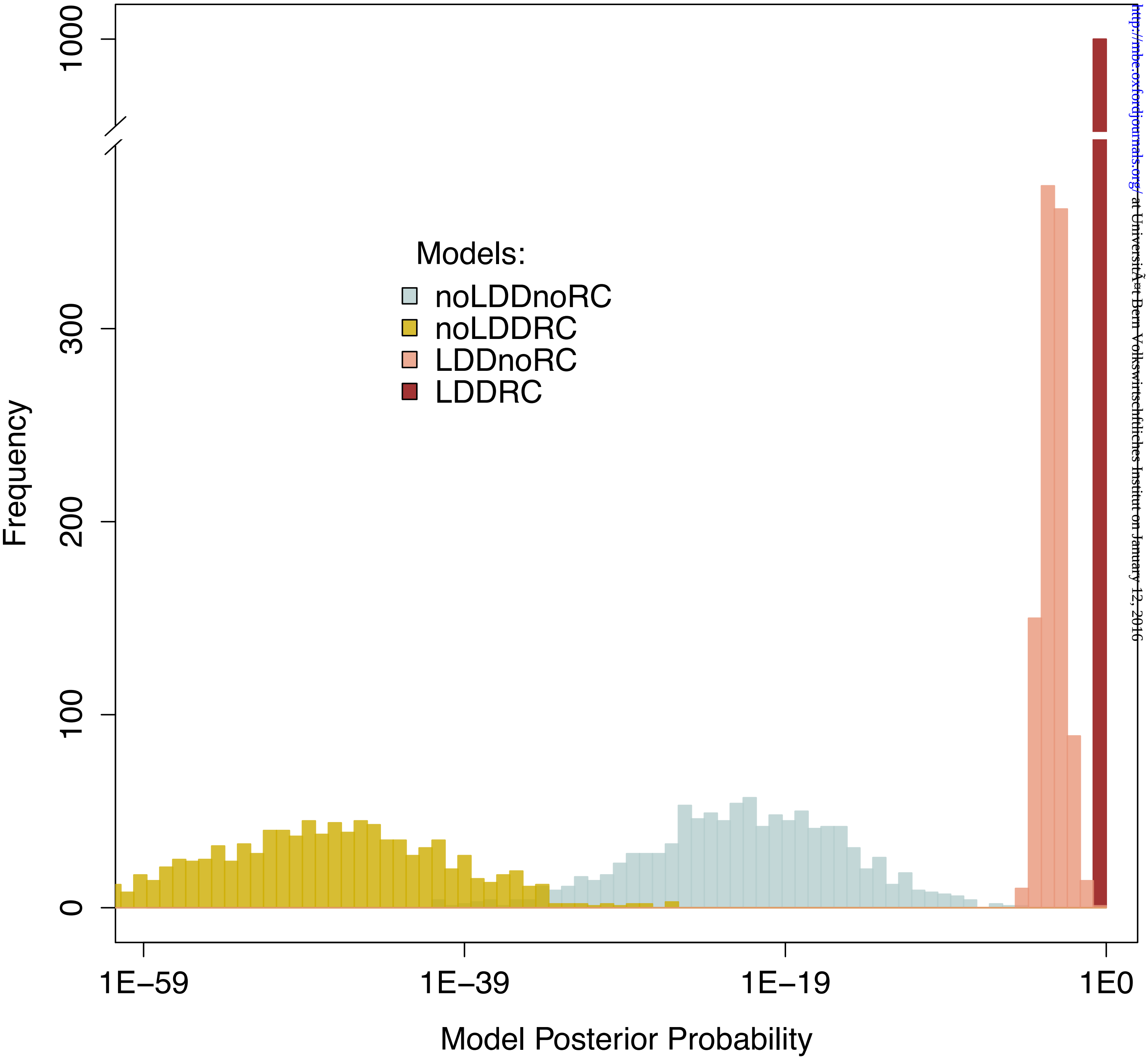
Figure 4. Distributions of the posterior probabilities of the six scenarios involving LDD (*LDDRC, LDDRCep*, *LDDRCop*, *LDDnoRC, LDDnoRCep*, *LDDnoRCop*) obtained over the 1,000 bootstrap datasets. Model posterior probabilities were computed using the multivariate logistic regression (Beaumont 2008) on the 10% best simulations (closest to the empirical data) among 20,000 simulations per evolutionary scenario.

FIG. 1



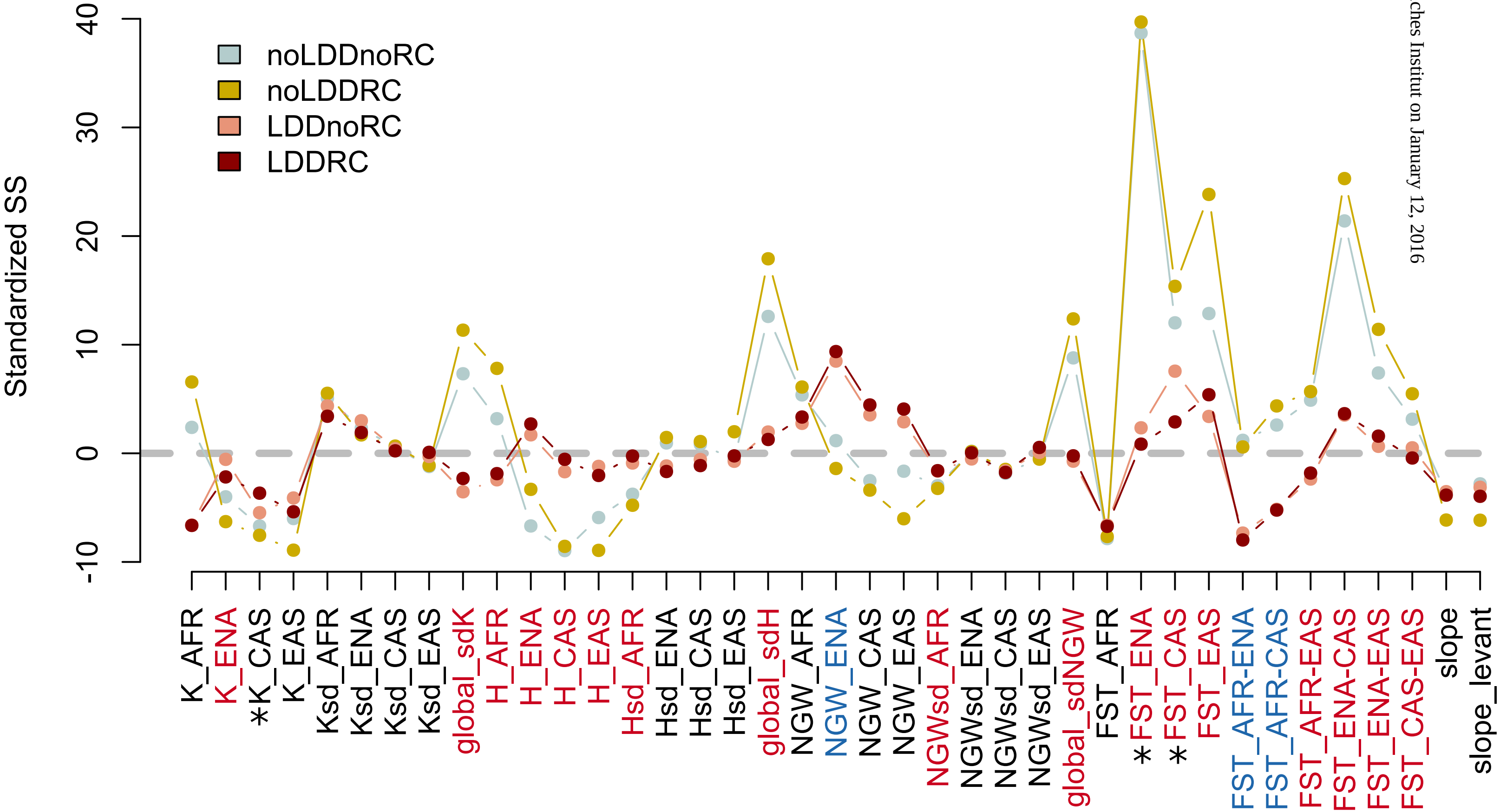
A)
B)
C)
D)
ENA
EAS
CAS
AFR
Orcadian (15)
Russian (25)
Basque (24)
Italian (21)
Adygei (17)
Uygur (10)
Yakut (25)
Mongola (10)
Japanese (29)
Burusho (25)
Han (44)
Mozabite (29)
Bedouin (46)
Hindi (28)
Makrani (25)
Assamese (25)
Mandinka (22)
Temani (21)
Tamil (29)
Dinka (17)
Bakola (42)
Sandawe (51)

FIG. 2



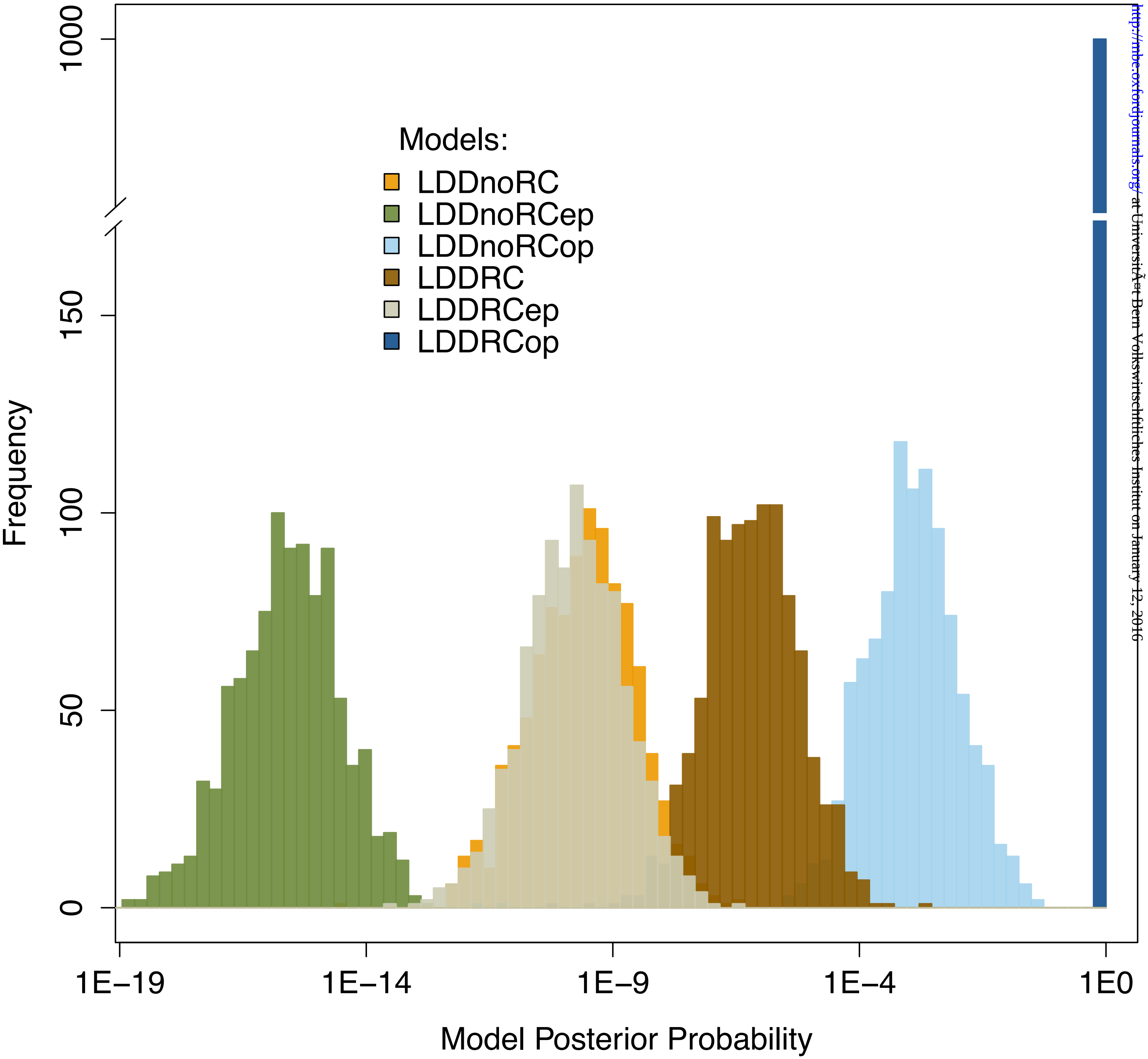
Models:
noLDDnoRC
noLDDRC
LDDnoRC
LDDRC
1000
300
200
100
0
Frequency
1E−59
1E−39
1E−19
1E0
Model Posterior Probability

FIG. 3



noLDDnoRC
noLDDRC
LDDnoRC
LDDRC
Standardized SS
40
30
20
10
0
-10
K_AFR
K_ENA
*K_CAS
K_EAS
Ksd_AFR
Ksd_ENA
Ksd_CAS
Ksd_EAS
global_sdK
H_AFR
H_ENA
H_CAS
H_EAS
Hsd_AFR
Hsd_ENA
Hsd_CAS
Hsd_EAS
global_sdH
NGW_AFR
NGW_ENA
NGW_CAS
NGW_EAS
NGWsd_AFR
NGWsd_ENA
NGWsd_CAS
NGWsd_EAS
global_sdNGW
FST_AFR
*FST_ENA
*FST_CAS
FST_EAS
FST_AFR-ENA
FST_AFR-CAS
FST_AFR-EAS
FST_ENA-CAS
FST_ENA-EAS
FST_CAS-EAS
slope
slope_levant

FIG. 4



Models:
LDDnoRC
LDDnoRCep
LDDnoRCop
LDDRC
LDDRCep
LDDRCop
Frequency
0
50
100
150
1000
1E−19
1E−14
1E−9
1E−4
1E0
Model Posterior Probability